Beyond CMOS computing

1. CMOS Scaling

Dmitri Nikonov

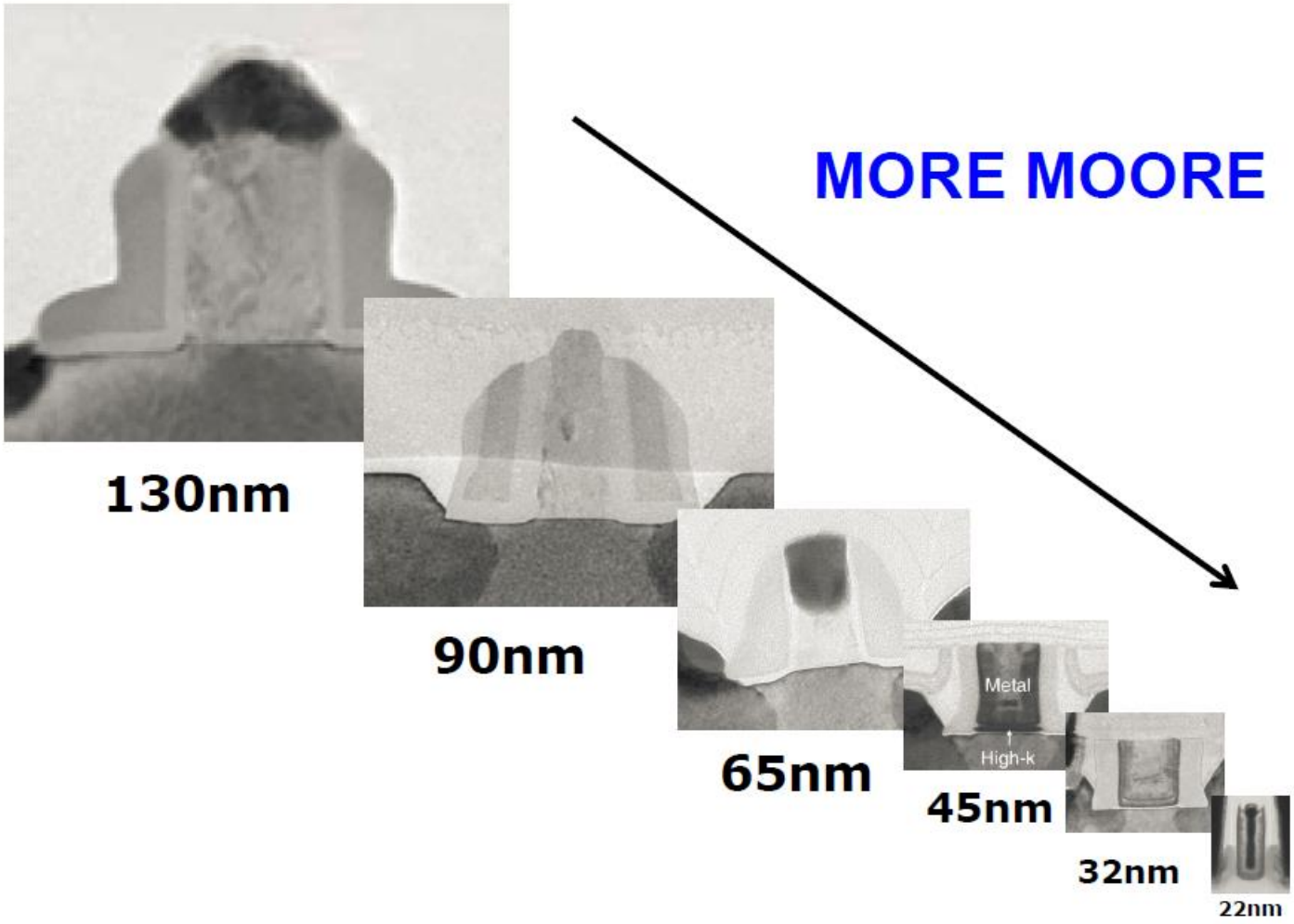
Thanks to Kelin Kuhn

Dmitri.e.nikonov@intel.com

Outline

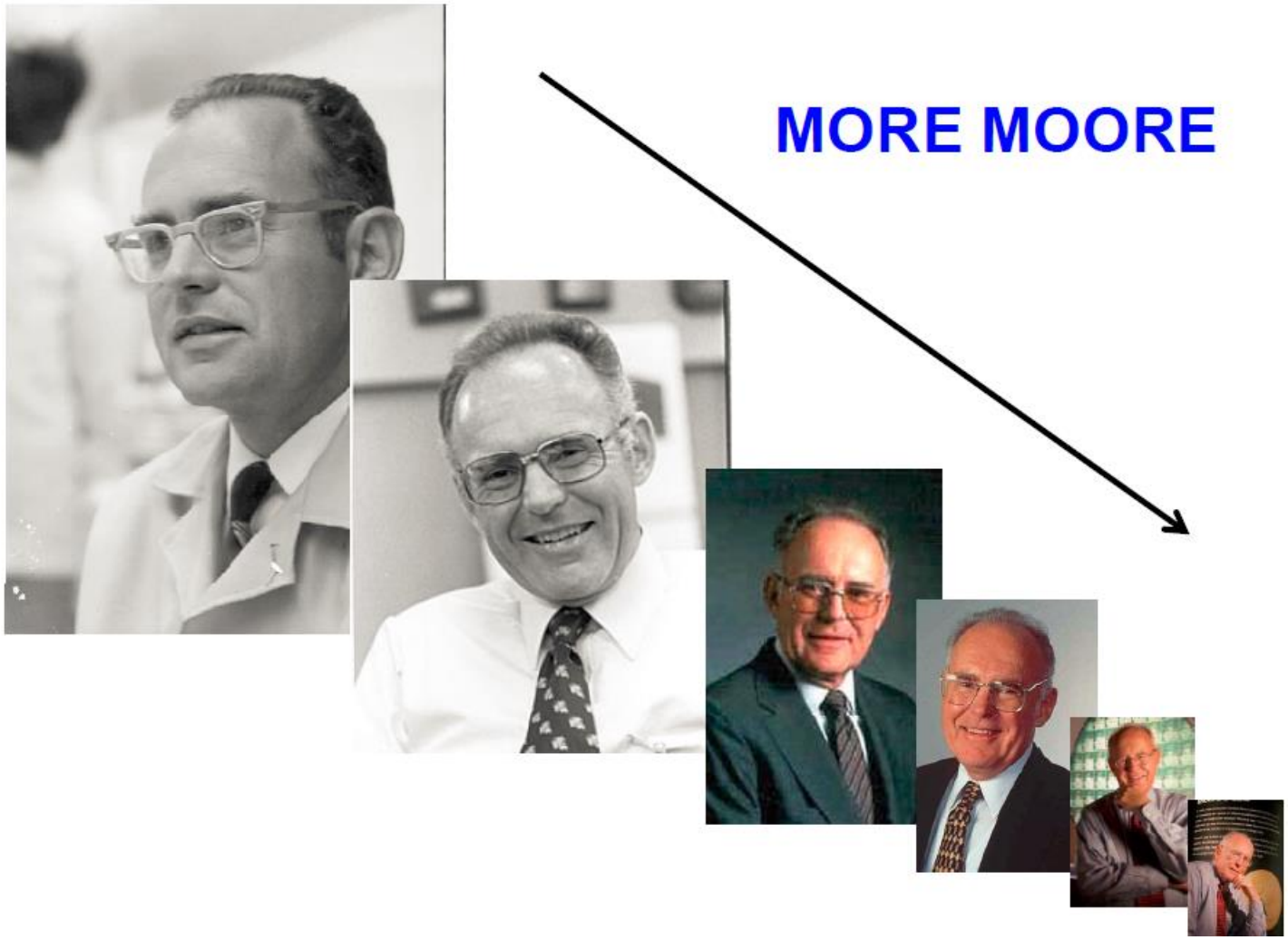
- Moore's law = scaling
- Performance improvement with scaling
- Latest: tri-gate transistors
- Fundamental limits to scaling

Moore's Law



Transistor size becoming smaller

Moore's Law



Gordon Moore becoming wiser

Scaling Falsifies Predictions

In the limit, microscope objectives with 0.95 N.A. are available and, provided very small fields (200μ x 200μ) are adequate, linewidths < 0.4μ should be achievable under carefully controlled laboratory conditions, and in very thin resist layers.

Depth of field will be reduced to about ± 0.2μ. Deep U.V. (λ = 200nm - 260nm) lenses will be difficult to build because of the lack of materials that are transparent at these wavelengths and yet have relatively high refractive indices.

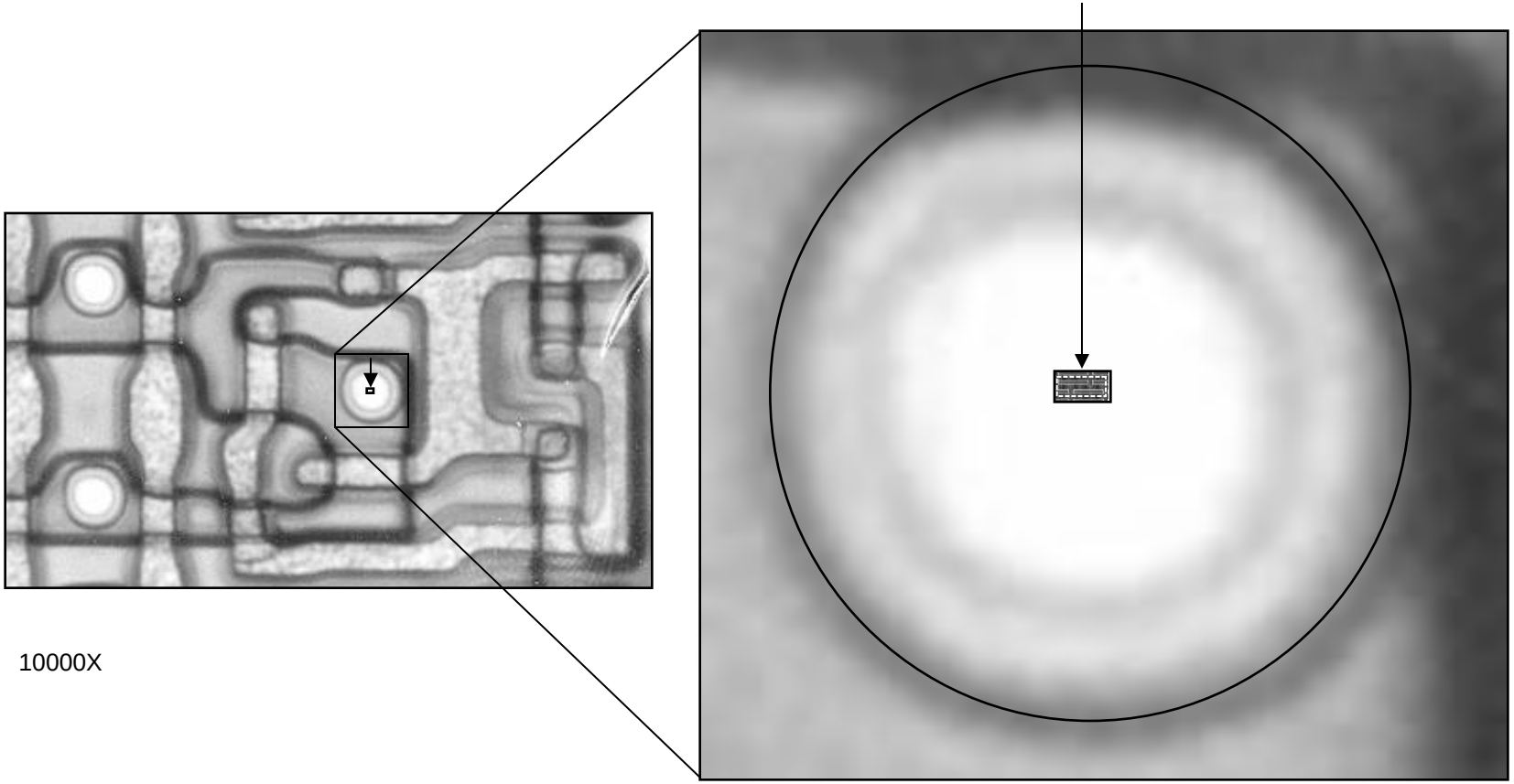
1980: Optical Lithography
Limit ~ 400nm

IEDM Plenary Session 1980 (Broers)

How Far Scaling Went

1980 SRAM Cell: 1700 um²

22nm SRAM Cell: 0.092 um²

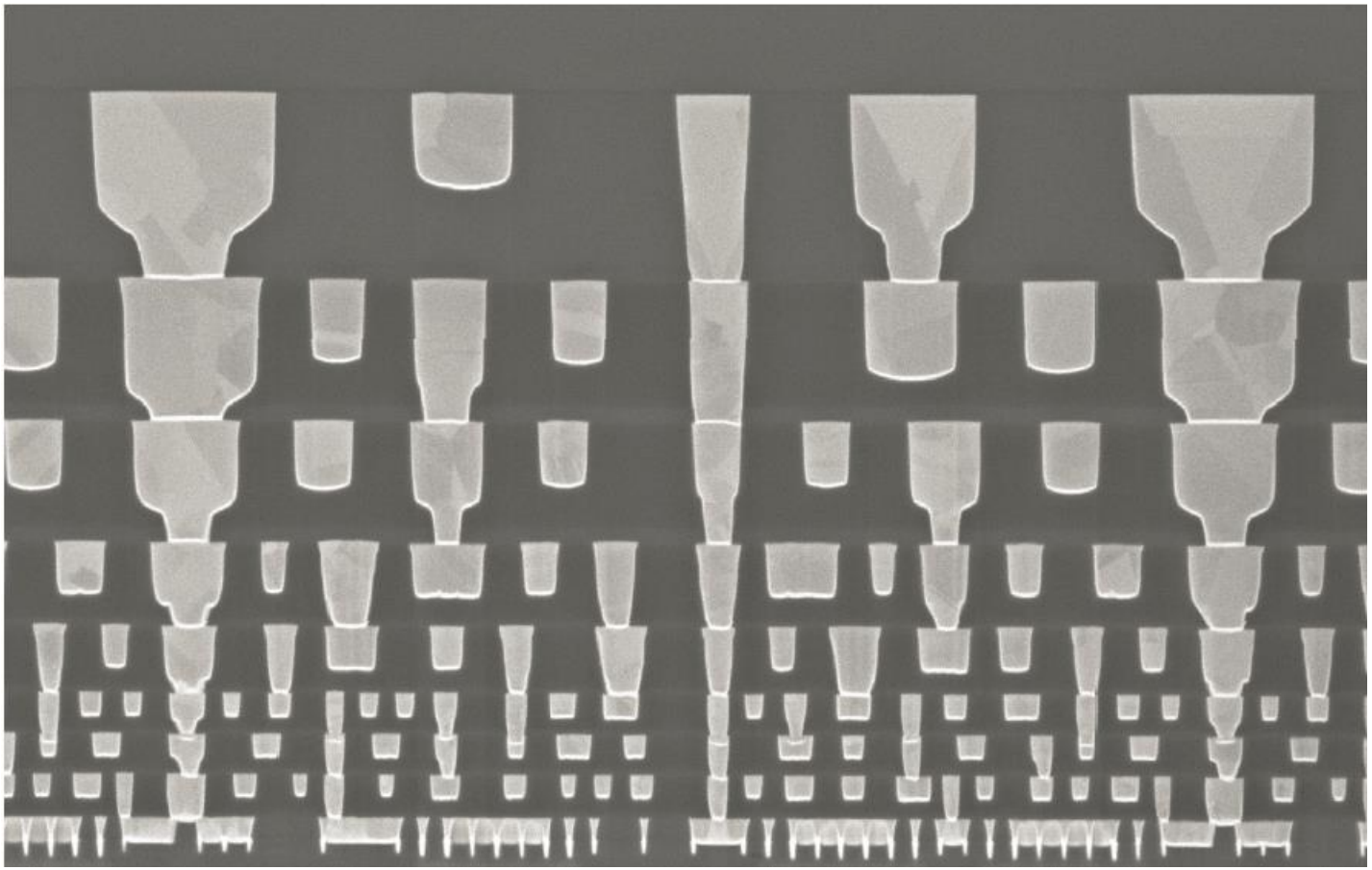


10000X

Small enough that a 2011 22nm SRAM cell is dwarfed by a 1980 SRAM cell CONTACT

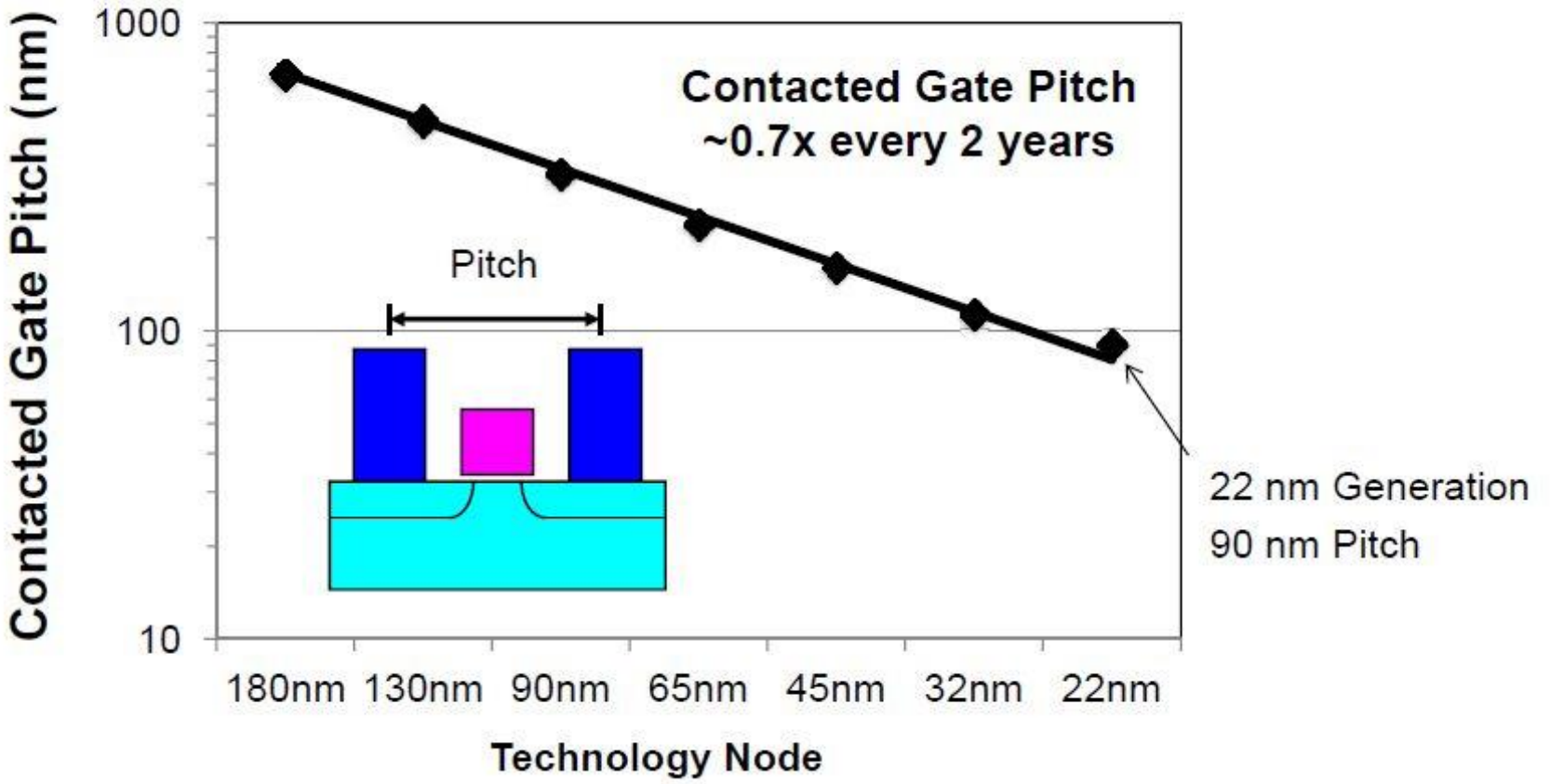
K. Kuhn, MIT invited seminar (MTL), 45nm High-k + Metal Gate Logic Technology, 5-19-08 (images from archives Mark Bohr, 2007)

Metal Interconnects



9 levels of metal

Contacted Gate Pitch

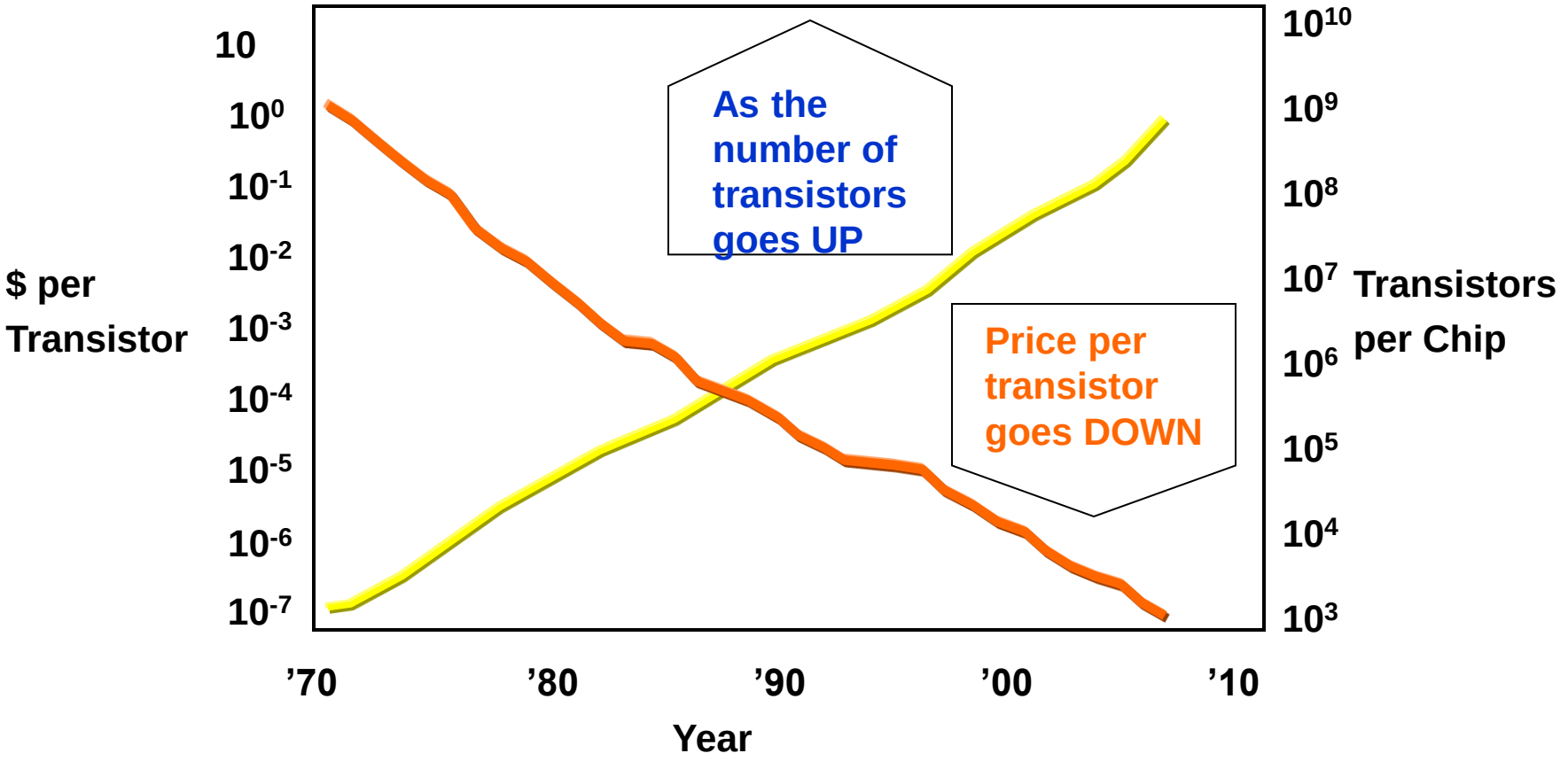


Transistor gate pitch continues to scale 0.7x every 2 years.

Proves to be 4*F. F is the label for process generations.

M. Bohr, ISCC, 2009.

Economics of Moore's Law



"Doubling of number of transistors per chip every 2 years".

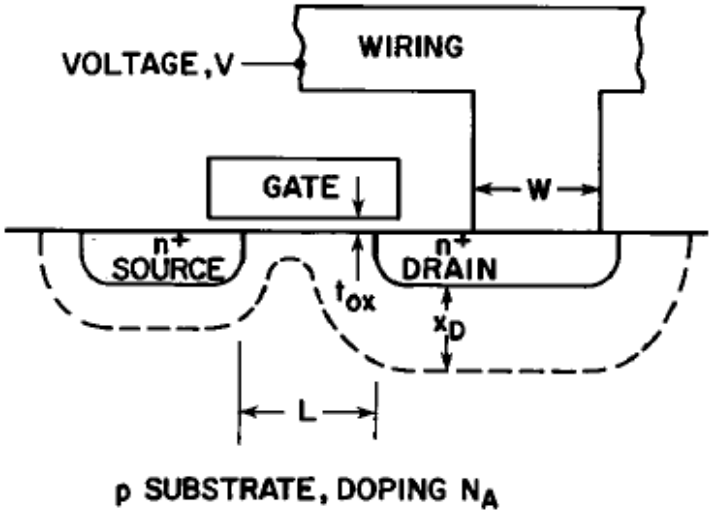
Lowers cost per transistor.

Self-fulfilling prophesy.

Original paper: G.E. Moore, Electronics 19, 114 (1965)

Classic Scaling

Device or Circuit Parameter	Scaling Factor
Device dimension *tox, L, W*	1/κ
Doping concentration *Na*	κ
Voltage *V*	1/κ
Current *I*	1/κ
Capacitance *εA/t*	1/κ
Delay time/circuit *VC/I*	1/κ
Power dissipation/circuit *VI*	$1/\kappa^2$
Power density *VI/A*	1

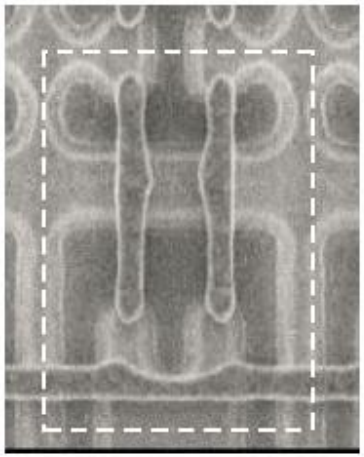


R. Dennard, IEEE JSSC, 1974

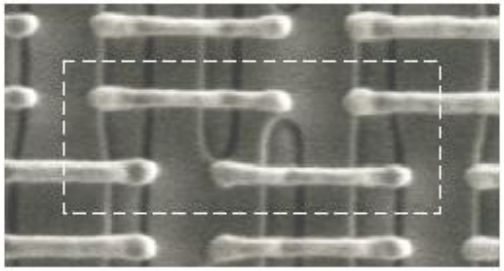
Classical MOSFET scaling was first described by Dennard in 1974

Dennard, IEEE JSSC, 1974

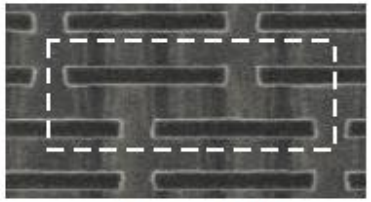
Process Evolution Over Time



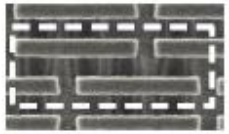
90nm – TALL
1.0 μm²



65nm – WIDE - 0.57 μm²



45nm – WIDE
0.346 μm²

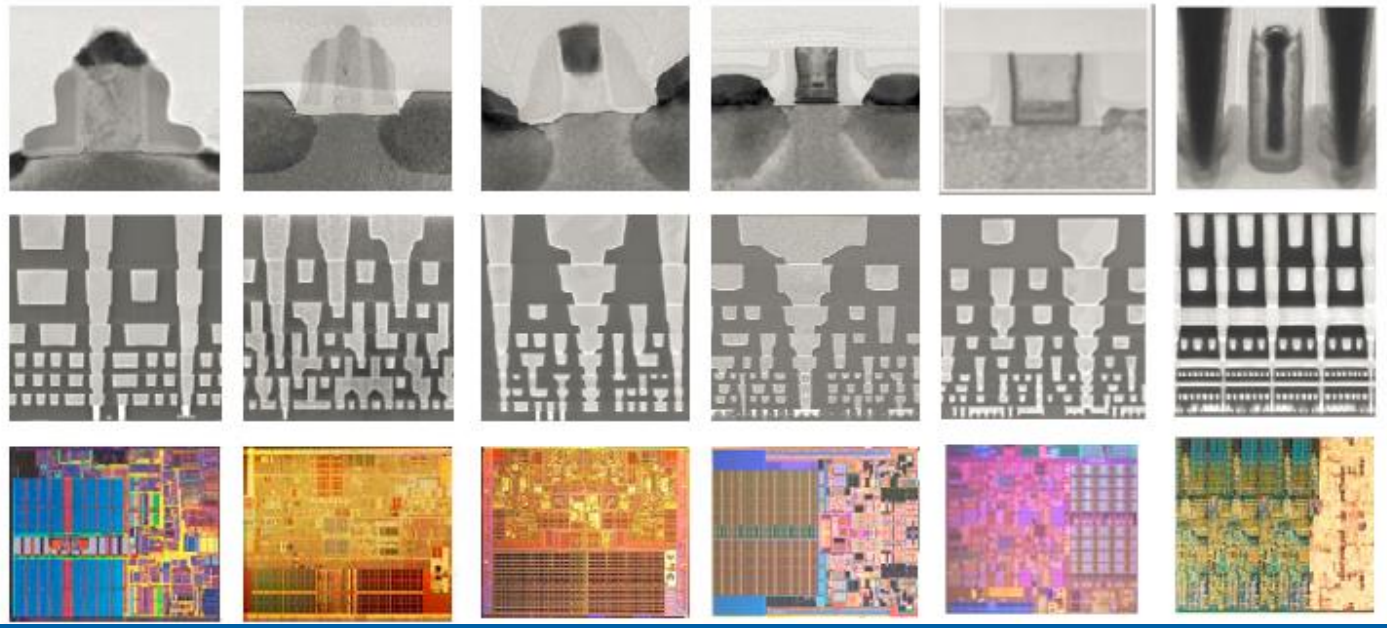


32nm – WIDE
0.171 μm²

22nm – WIDE
0.092 μm²

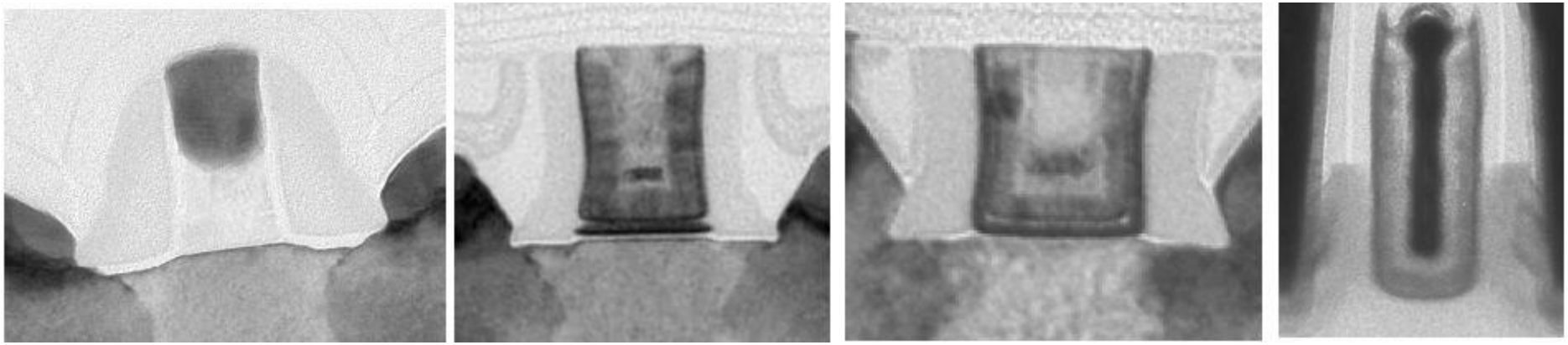
130 nm	90 nm	65 nm	45 nm	32 nm	22 nm
2001	2003	2005	2007	2009	2011

Bohr Intel Press release 2012



Inflection in Scaling

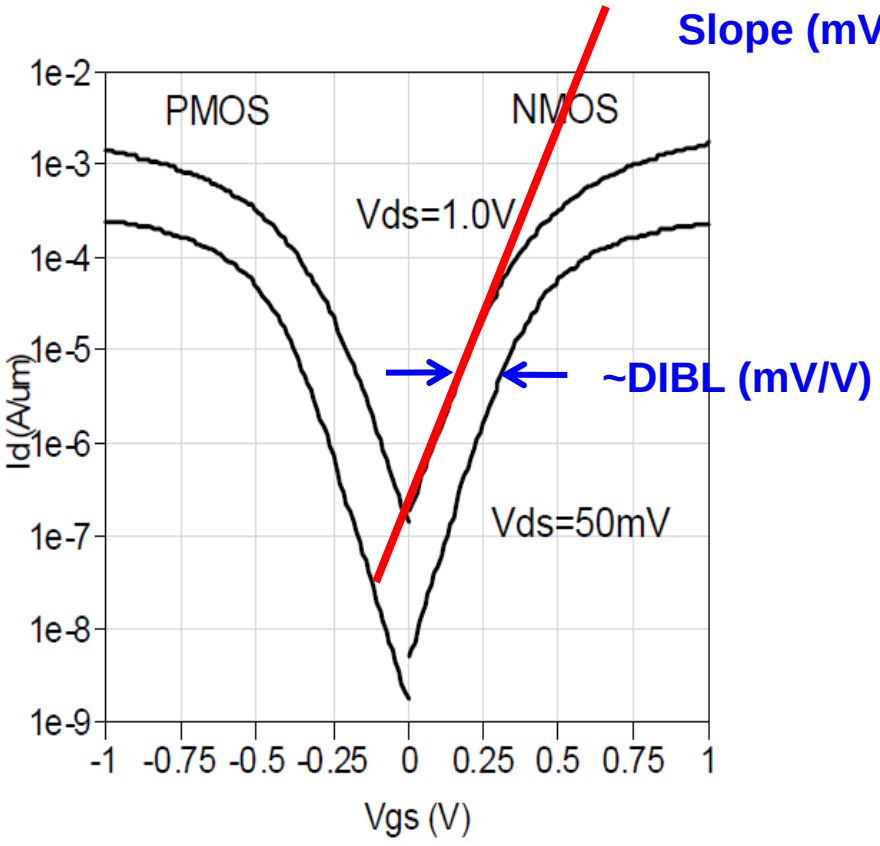
THEN	NOW
Scaling drove down cost	Scaling drives down cost
Scaling drove performance	Materials drive performance
Performance constrained	Power constrained
Active power dominates	Standby power dominates
Independent design-process	Collaborative design-process



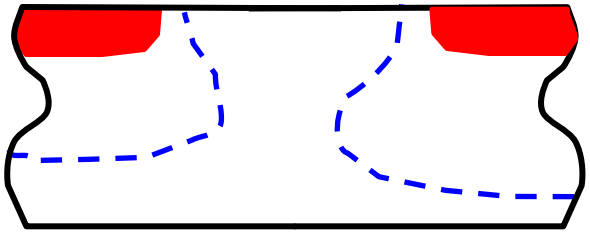
65nm 45nm 32nm 22nm

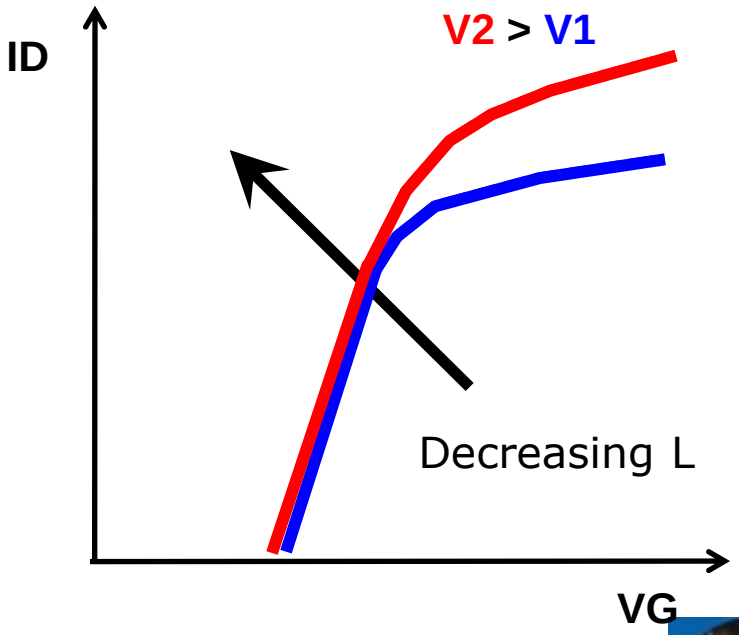
Images from Bai, Mistry, Natarajan, Auth IEDM/VLSI, 2004/7/8/12
(see course required reading)

Short Channel Effects (SCE)

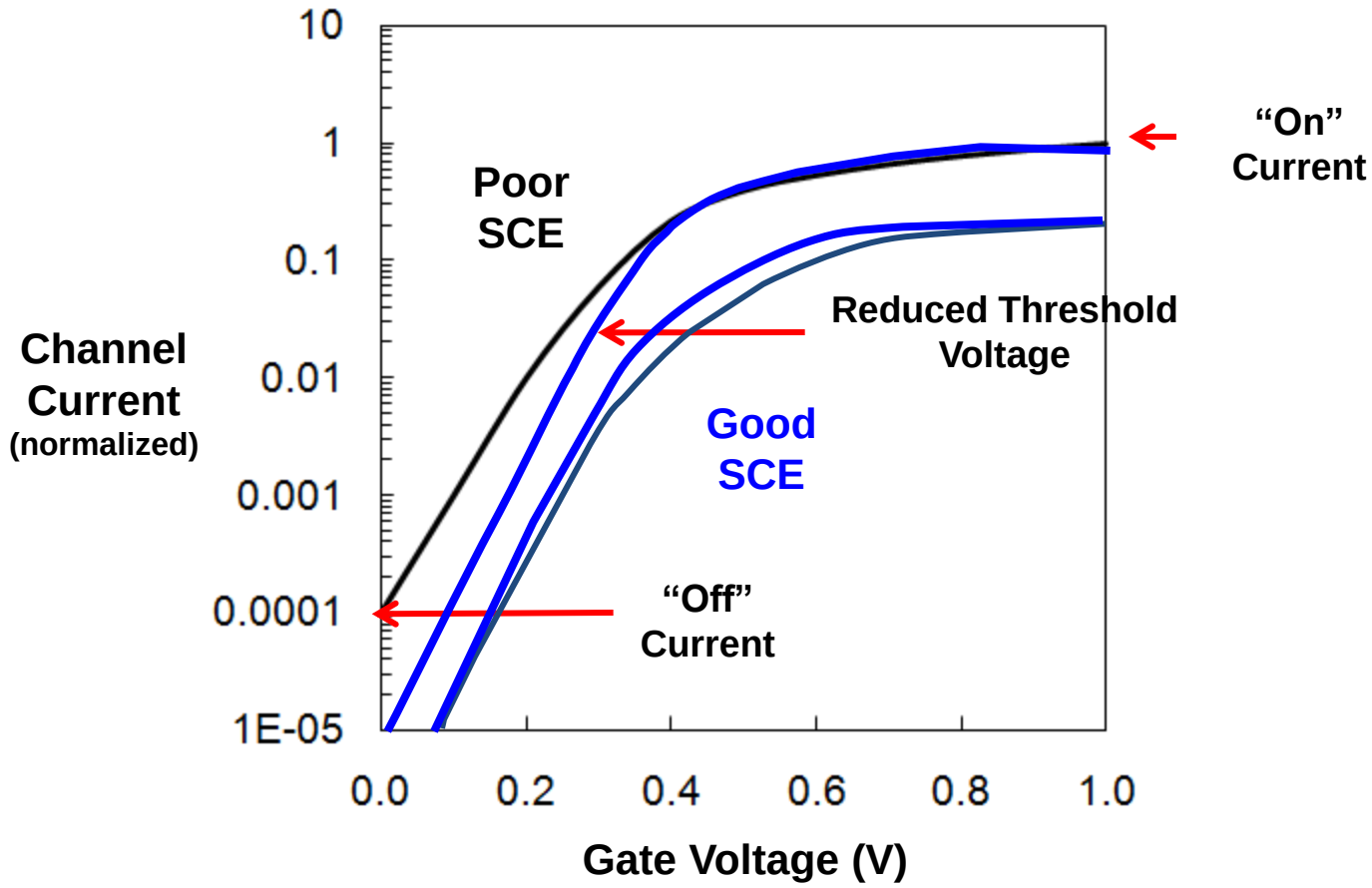


Degradation of short channel effects



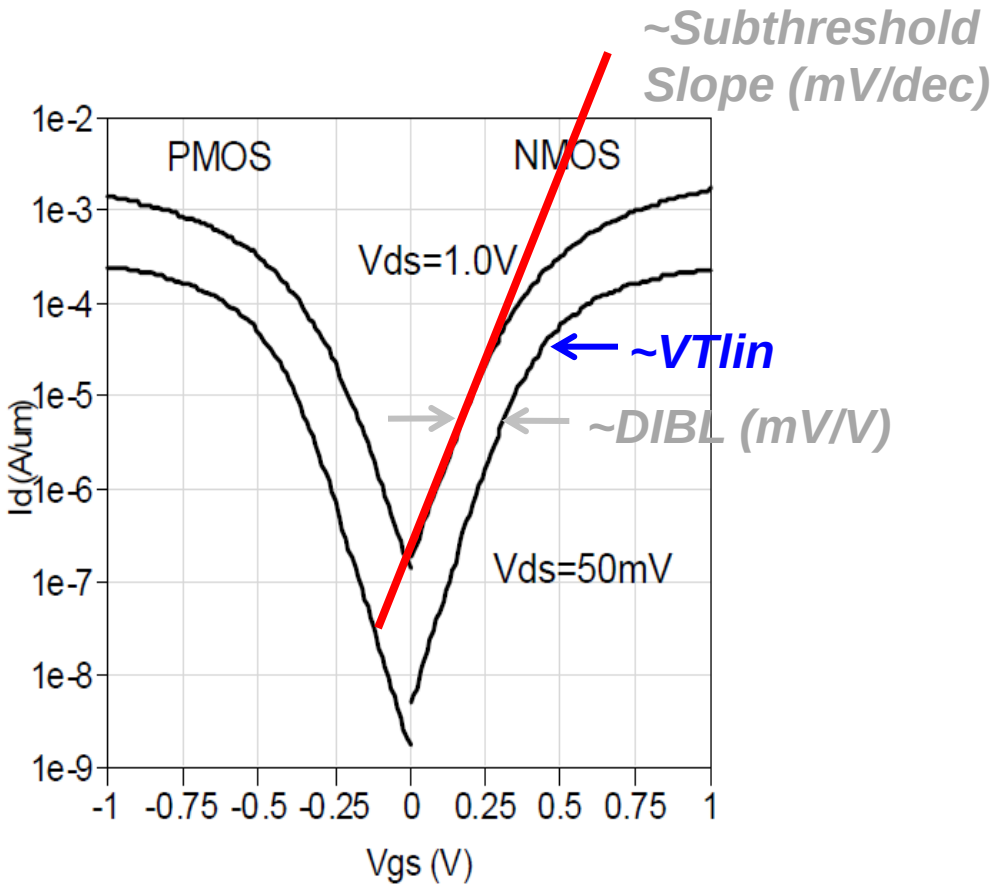


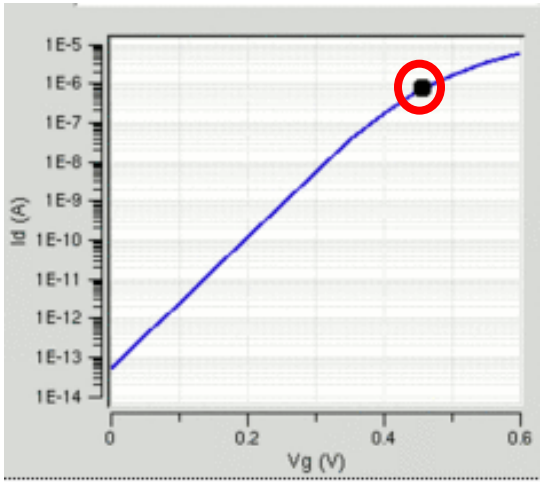
Electrostatics Benefits

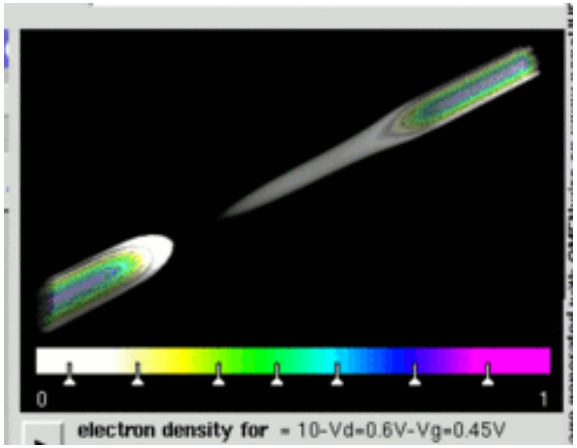


Basic ID-VG

Threshold Voltage

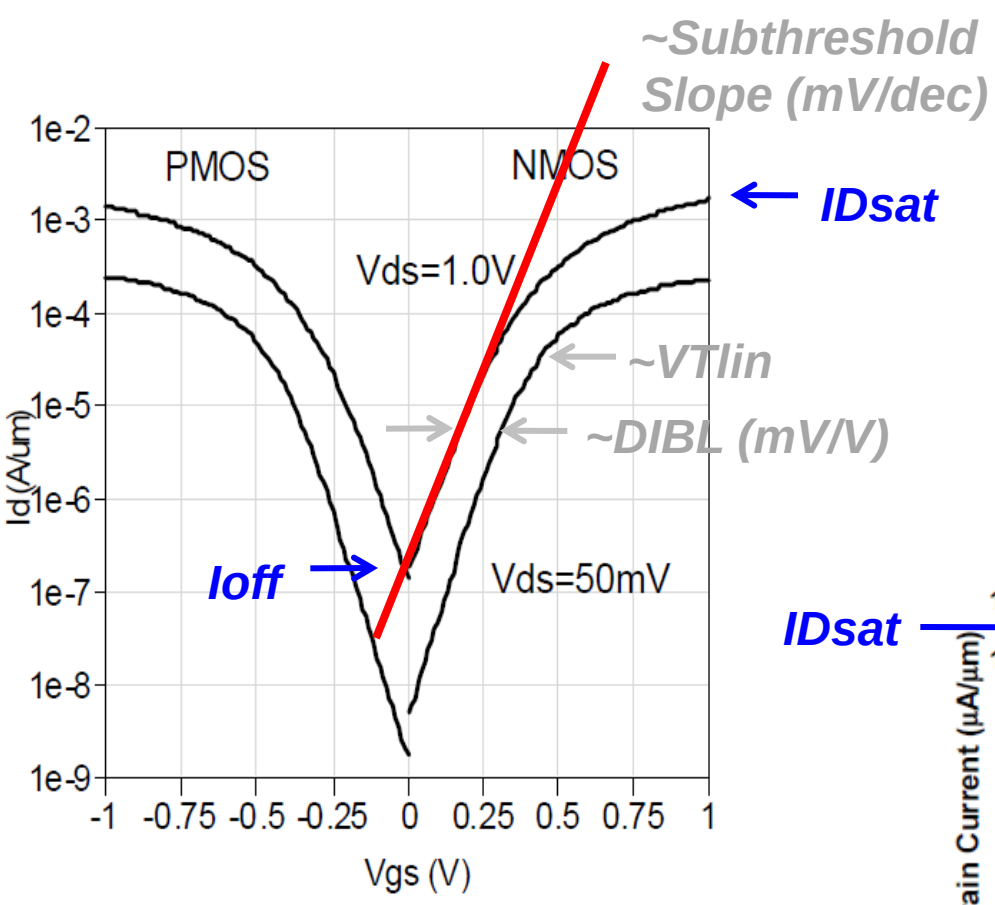


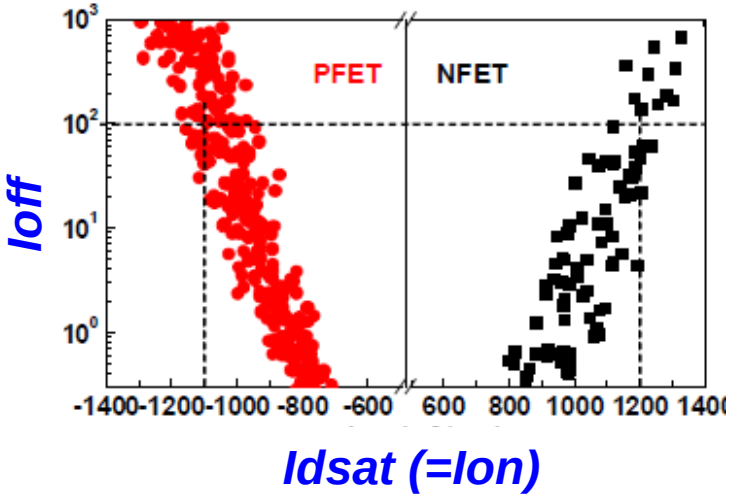


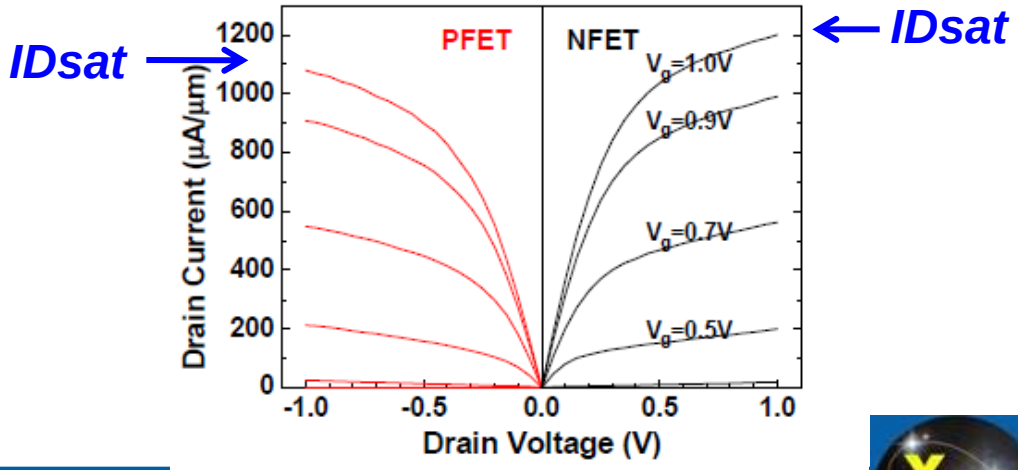


http://en.wikipedia.org/wiki/Threshold_voltage

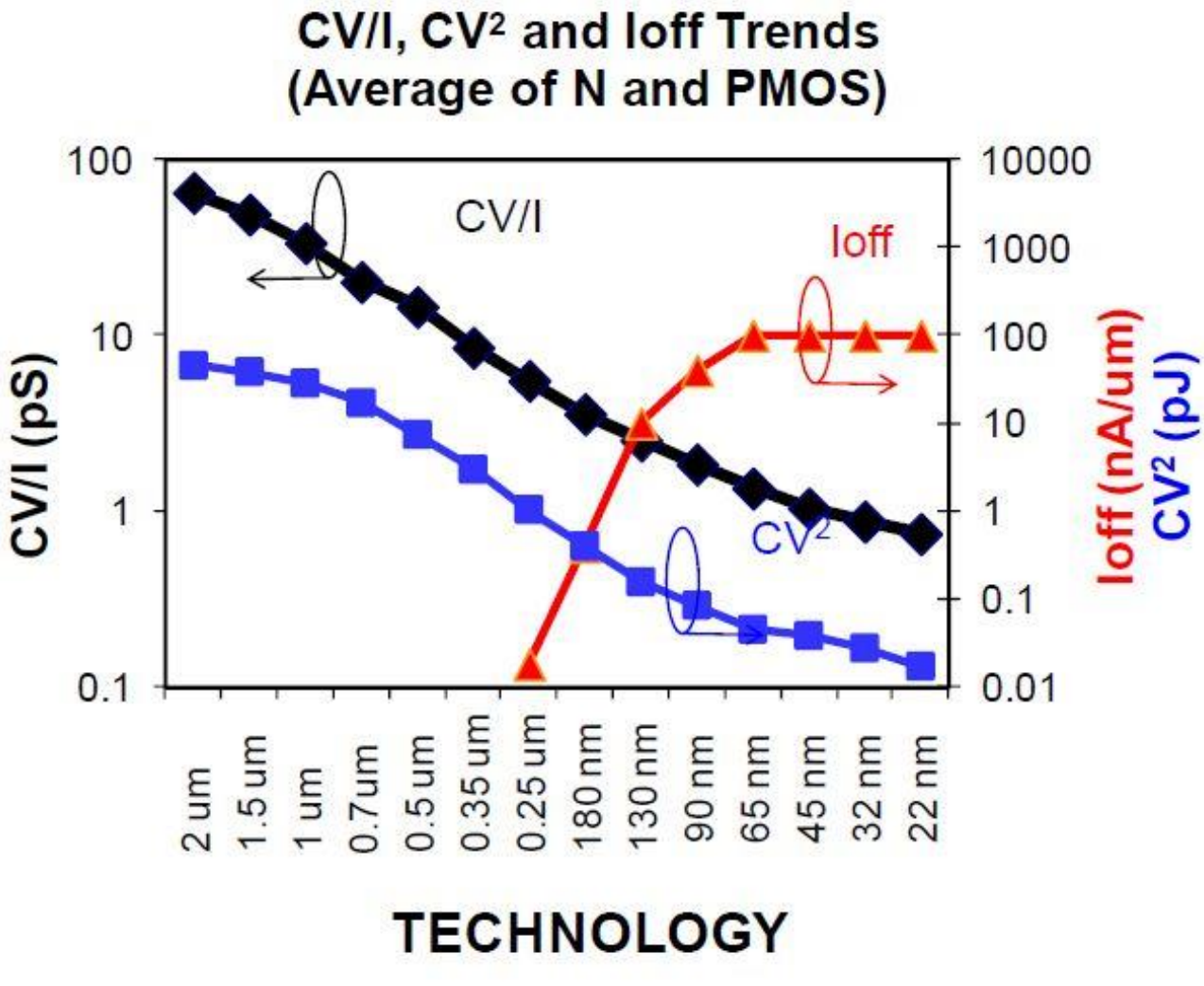
On Current, Off Current





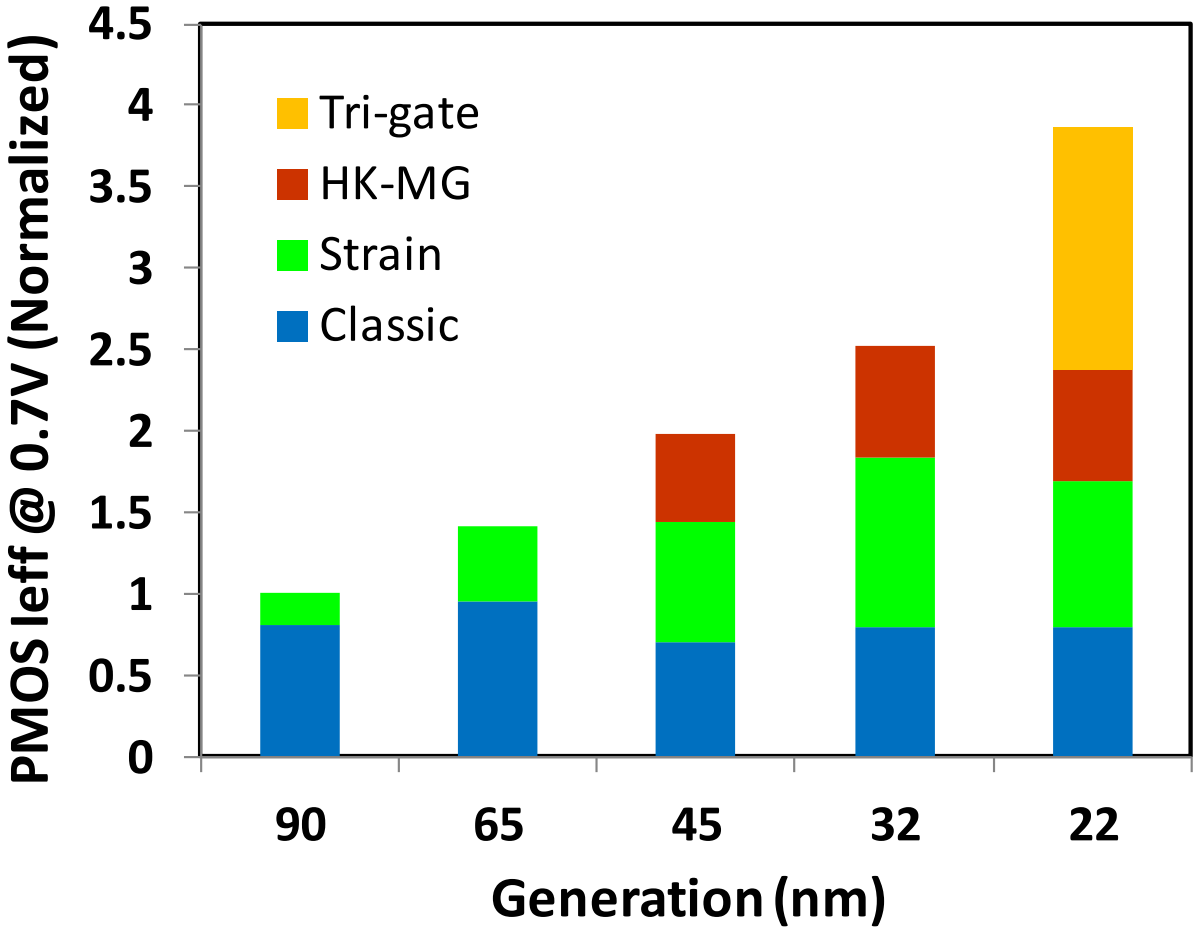


Performance from Scaling



Natarajan, Intel, IEDM, 2008

Where Performance Comes From

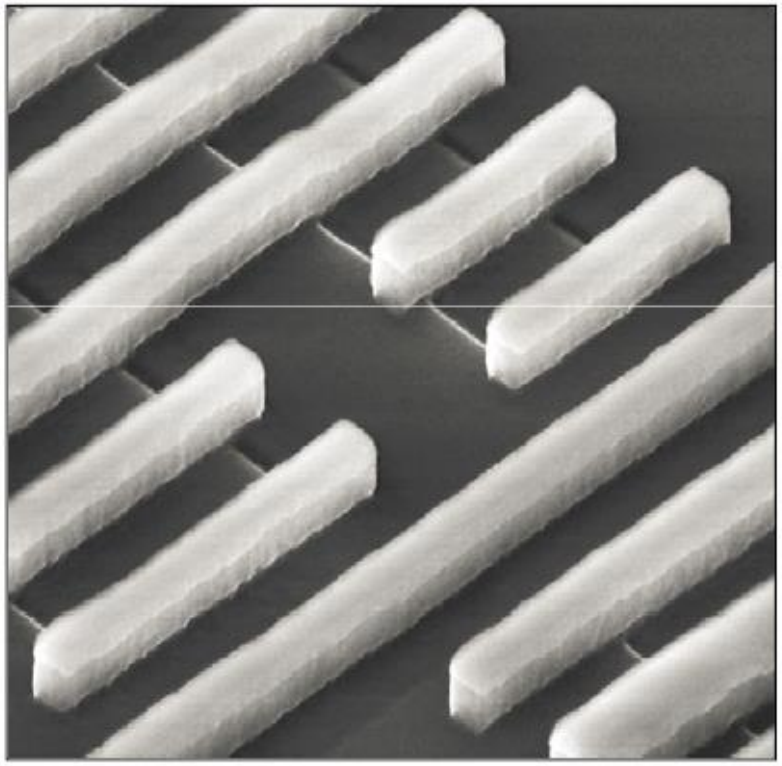


Strain and High-k + metal gate are key enablers past the 90nm node

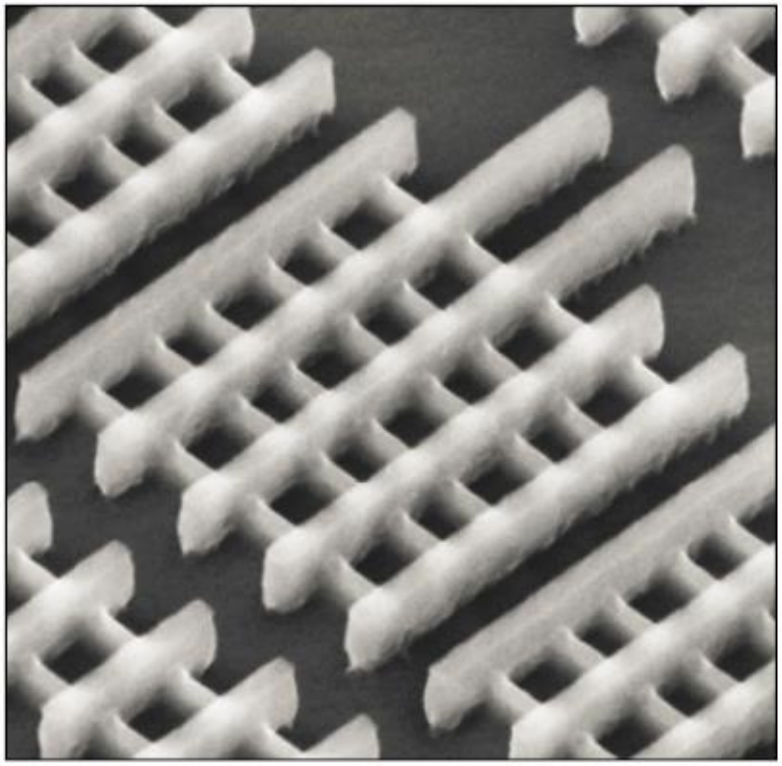
K.J. Kuhn, Moore's Law past 32nm: Future Challenges in Device Scaling, Solid State Devices and materials conference, plenary, October 2009

Intel Announced First Tri-Gate (2011)

32 nm Planar Transistors

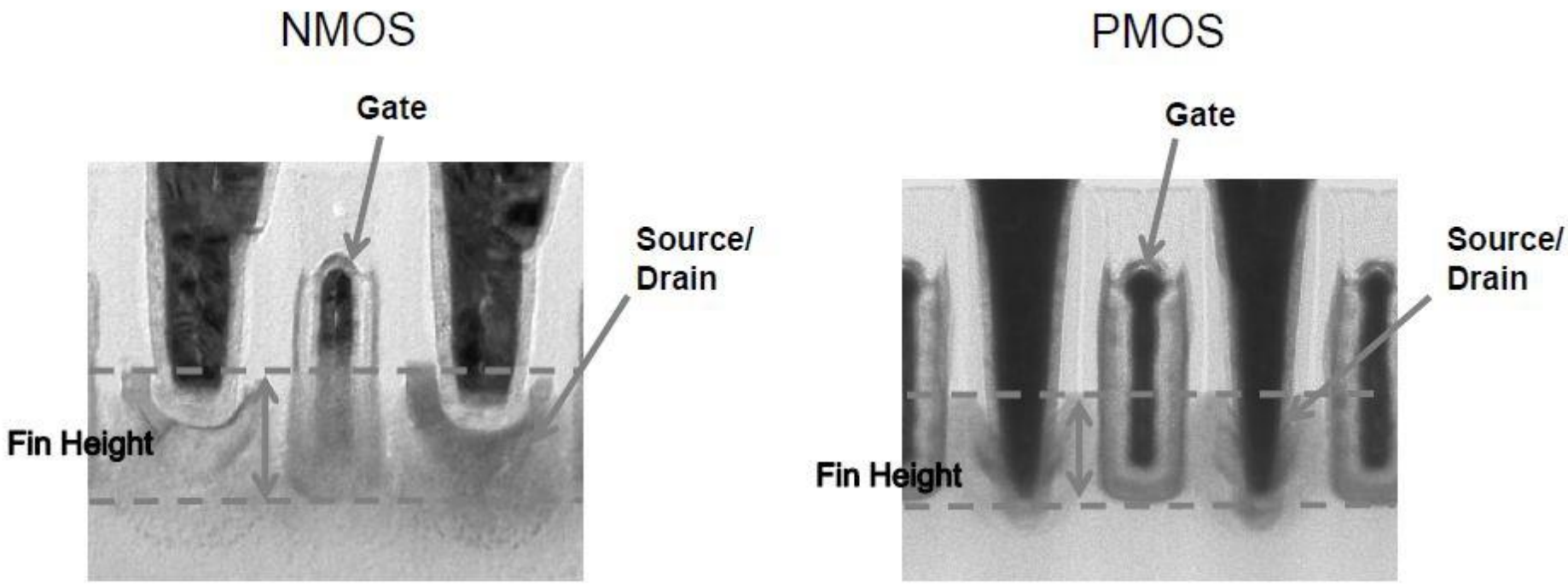


22 nm Tri-Gate Transistors



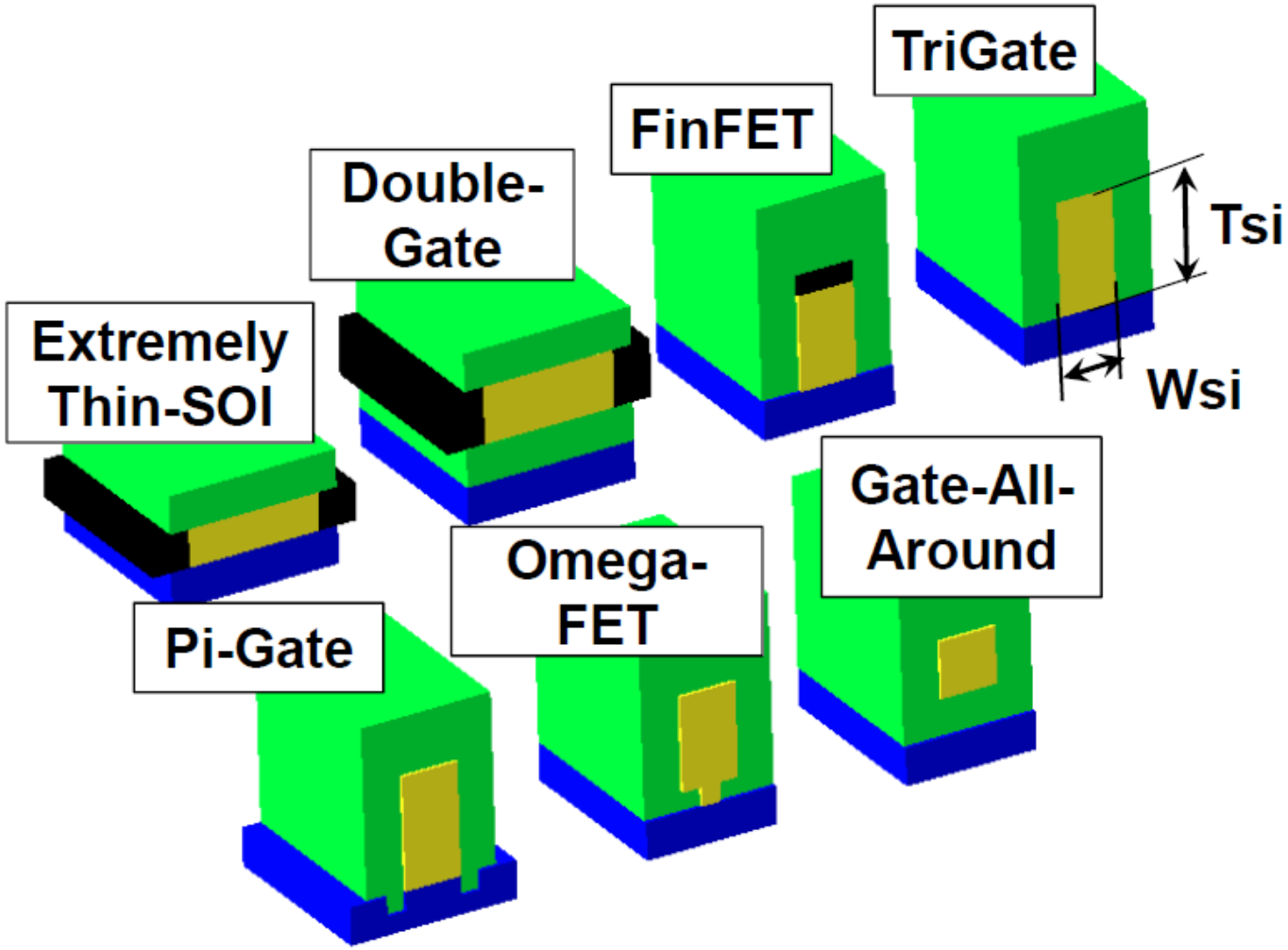
Mark Bohr, Kaizad Mistry: Intel, April 25th, press release

Close-Up on Tri-gate Transistors

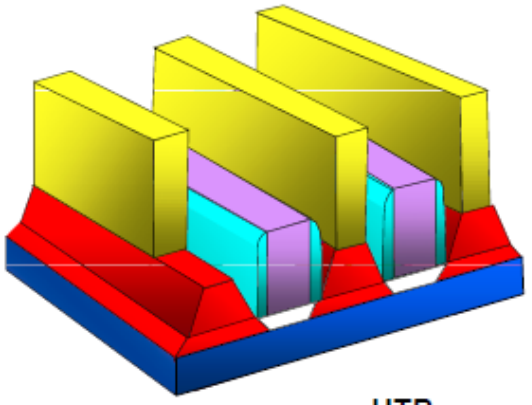


The 22nm process technology was the first to exploit fin-based Tri-Gate devices and combine their benefits with strained silicon and high-k/metal-gate

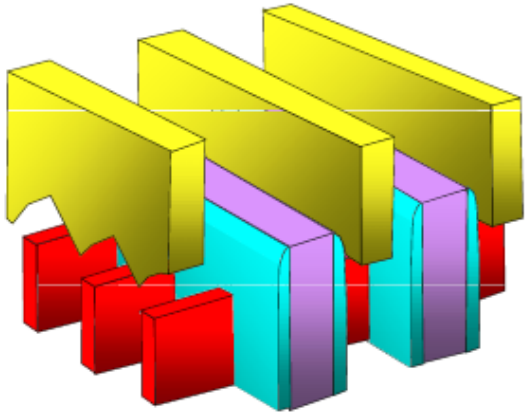
Nomenclature of Non-Planar Devices



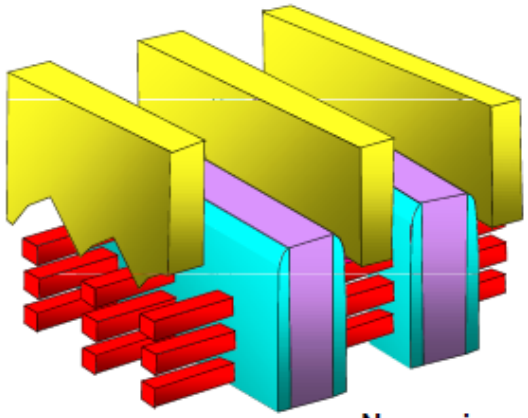
Where Non-Planar Can Go?



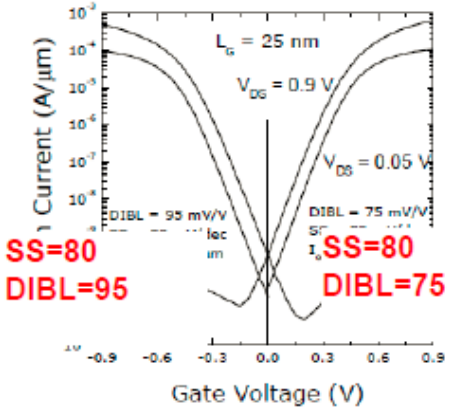
UTB



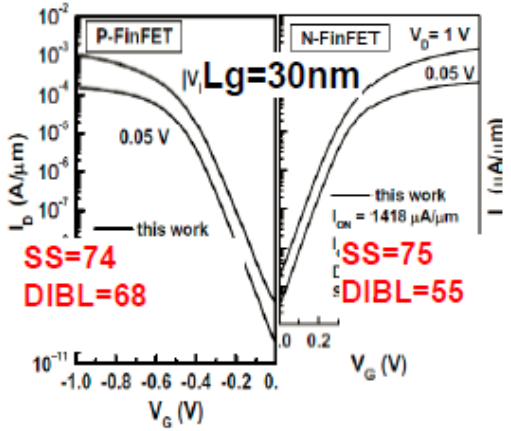
FinFET/Trigate



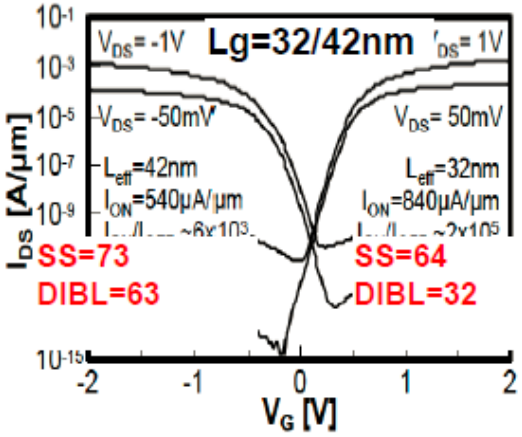
Nanowire



Cheng – IEDM 2009 (IBM)

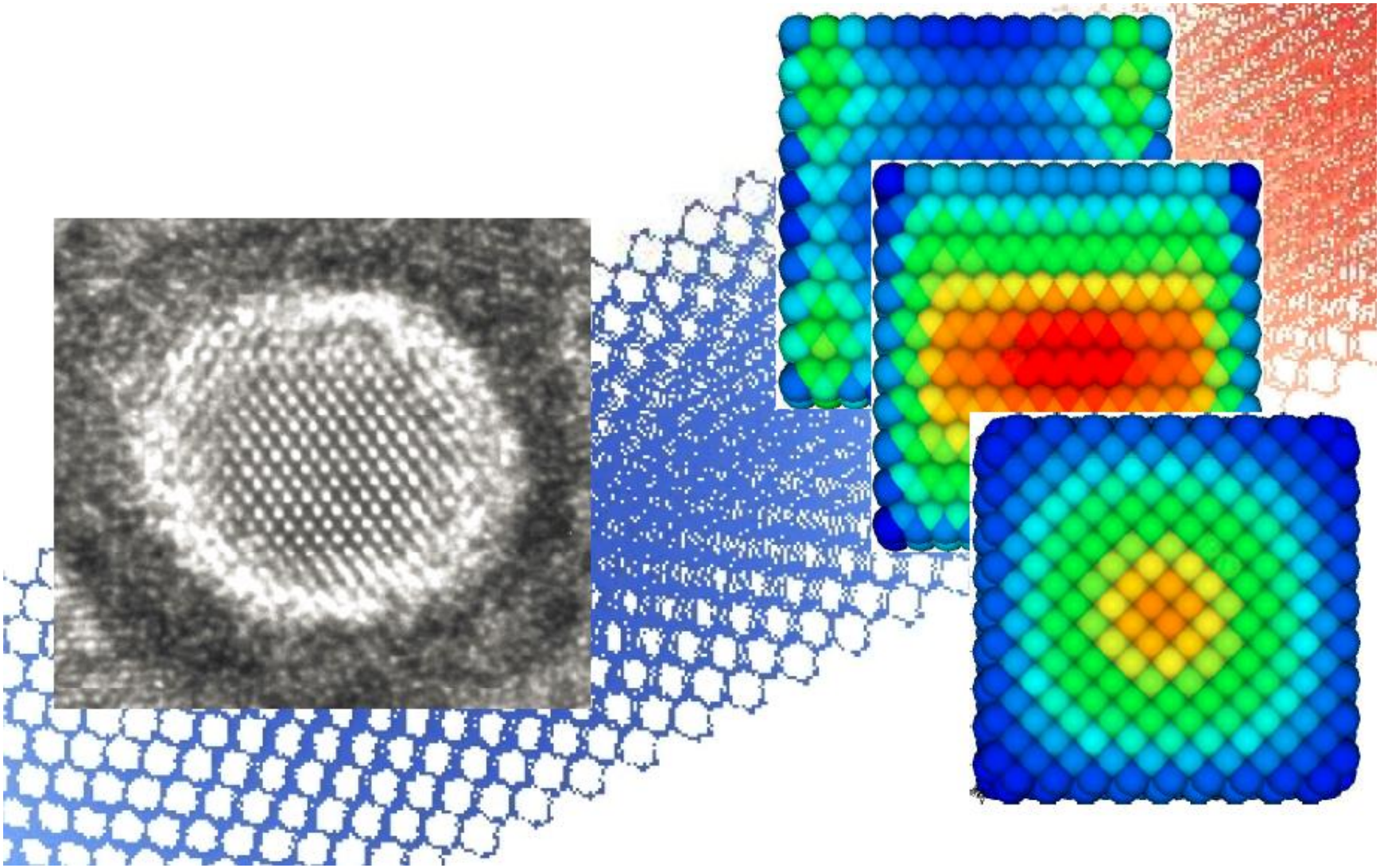


Yeh – IEDM 2010 (TSMC)



Tachi – IEDM 2010 (CEA-LETI)

Working With Atomic Dimensions

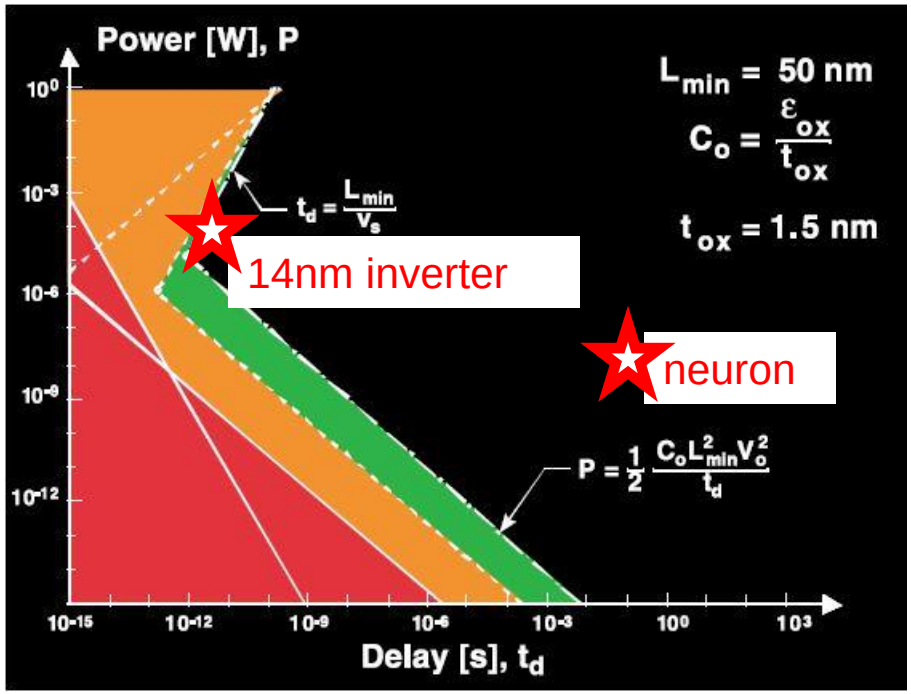


International Technology Roadmap for Semiconductors

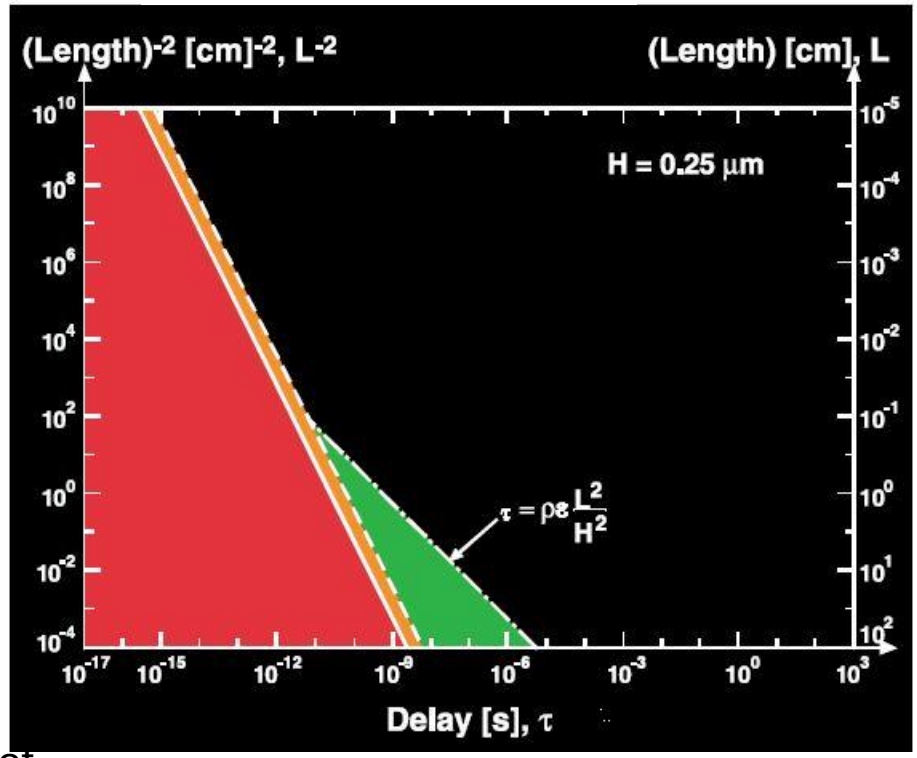
Year of Production	*2011*	*2012*	*2013*	*2014*	*2015*	*2016*	*2017*	*2018*	*2019*	*2020*	*2021*	*2022*	*2023*	*2024*	*2025*	*2026*
MPU/ASIC Metal 1 (M1) ½ Pitch (nm) (contacted)	*38*	*32*	*27*	*24*	*21*	*18.9*	*16.9*	*15.0*	*13.4*	*11.9*	*10.6*	*9.5*	*8.4*	*7.5*	*6.7*	*6.0*
L_g: Physical Lgate for HP Logic (nm) [1]	***24***	***22***	***20***	***18***	***17***	***15.3***	***14.0***	***12.8***	***11.7***	***10.6***	***9.7***	***8.9***	***8.1***	***7.4***	***6.6***	***5.9***
V_{dd}: Power Supply Voltage (V) [2]																
Bulk/FD SOI/MG	0.90	0.87	0.85	0.82	0.80	0.77	0.75	0.73	0.71	0.68	0.66	0.64	0.62	0.61	0.59	0.57
EOT: Equivalent Oxide Thickness (nm) [3]																
Extended Planar Bulk	0.88	0.84	0.79	0.73	0.67	0.61	0.55									
FD SOI			0.84	0.8	0.76	0.72	0.68	0.63	0.58	0.54						
MG					0.8	0.76	0.72	0.68	0.65	0.62	0.59	0.56	0.53	0.5	0.47	0.45
$I_{d,sat}$: NMOS Drive Current (μA/μm) [14]																
Extended Planar Bulk	1,320	1,367	1,422	1,496	1,582	1,670	1,775									
FD SOI			1,475	1,530	1,591	1,654	1,717	1,791	1,847	1,942						
MG					1,628	1,685	1,744	1,805	1,858	1,916	1,976	2,030	2,087	2,152	2,228	2,308
Equivalent Injection Velocity, v_{inj} (10^7 cm/s) [15]																
Extended Planar Bulk	1.03	1.09	1.11	1.18	1.24	1.33	1.39									
FD SOI			1.29	1.37	1.42	1.51	1.57	1.63	1.67	1.83						
MG					1.59	1.68	1.74	1.82	1.89	2.05	2.14	2.26	2.34	2.38	2.50	2.67
C_g Fringing Capacitance (fF/μm) [16]																
Extended Planar Bulk	0.24	0.24	0.24	0.24	0.24	0.24	0.24									
FD SOI			0.17	0.17	0.17	0.17	0.17	0.17	0.17	0.17						
MG					0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18	0.18
$C_{g,total}$: Total Gate Capacitance for Calculation of CV/I (fF/μm) [17]																
Extended Planar Bulk	0.936	0.898	0.872	0.851	0.834	0.816	0.805									
FD SOI			0.732	0.699	0.670	0.641	0.617	0.599	0.582	0.563						
MG					0.663	0.635	0.611	0.589	0.565	0.542	0.518	0.500	0.481	0.464	0.441	0.418
CV^2: NMOSFET Dynamic Power Indicator (fJ/μm) [18]																
Extended Planar Bulk	0.76	0.68	0.63	0.57	0.53	0.49	0.45									
FD SOI			0.53	0.47	0.43	0.38	0.35	0.32	0.29	0.26						
MG					0.42	0.38	0.34	0.31	0.28	0.25	0.23	0.21	0.19	0.17	0.15	0.14
τ =CV/I: NMOSFET Intrinsic Delay (ps) [19]																
Extended Planar Bulk	0.64	0.57	0.52	0.47	0.42	0.38	0.34									
FD SOI			0.42	0.38	0.34	0.30	0.27	0.24	0.22	0.20						
MG					0.32	0.29	0.26	0.24	0.21	0.19	0.17	0.16	0.14	0.13	0.12	0.10

Limits of Computing

Transistor Limits



Interconnect Limits



Moving from fundamental limits given by the laws of physics to practical limitations. These limits tend to be broken.

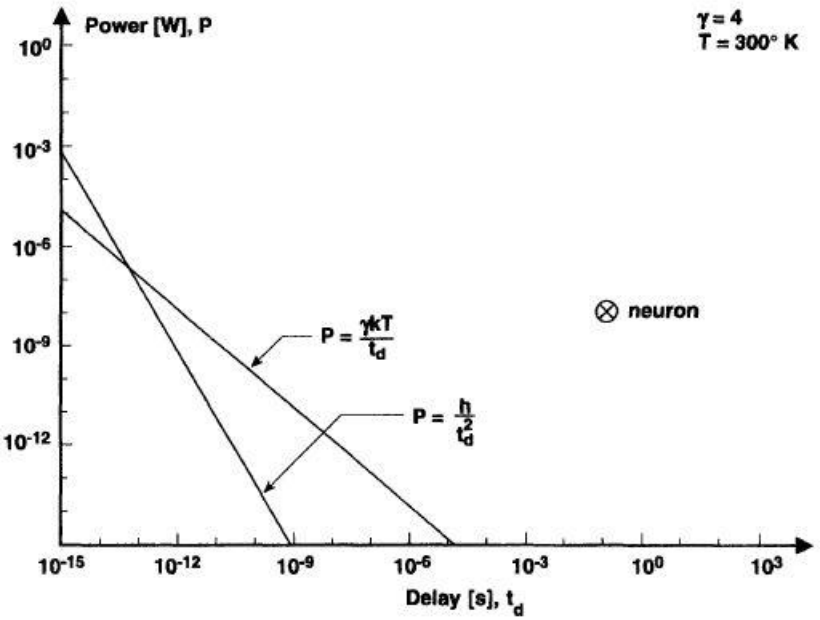
Meindl, Proc. IEEE 83, 619 (1995).
Meindl, Chen, Davis, Science 293, 2044 (2001).
Meindl, J. Vac. Sci. Technol. B 14(1), 192 (1996).

Fundamental limits

Material limits

Device limits

Equations for Fundamental Limits



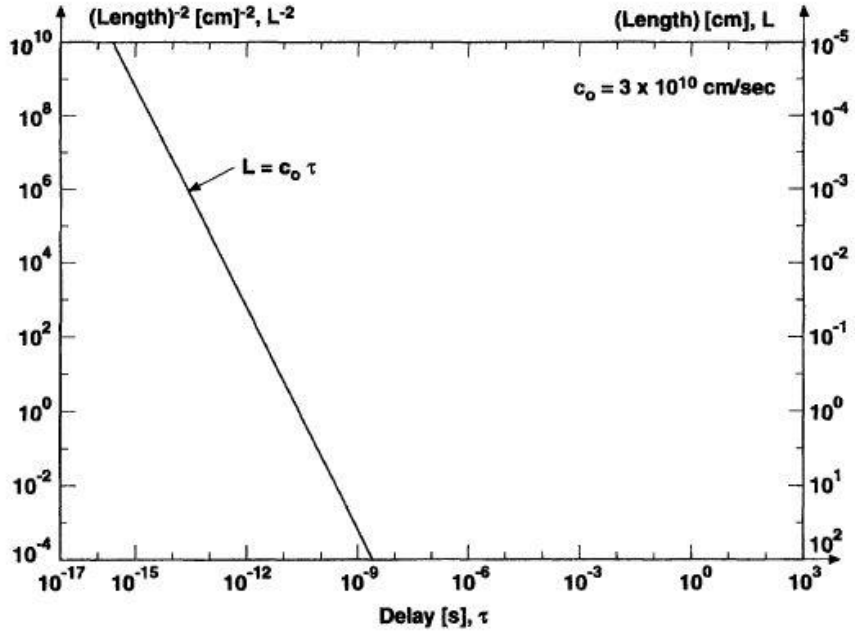
Thermodynamics: if energy is less than thermal – bit errors

$$E_{sw} = 4kT$$

Quantum Mechanics: energy time uncertainty

$$E_{sw} t_{sw} \geq h$$

Meindl, Proc. IEEE 83, 619 (1995).



Relativity: Signal no faster than the speed of light

$$L / \tau \leq c_0$$

Better Fundamental Limits

Thermodynamic limit on bit error ratio

$$E_{sw} = 3kT$$

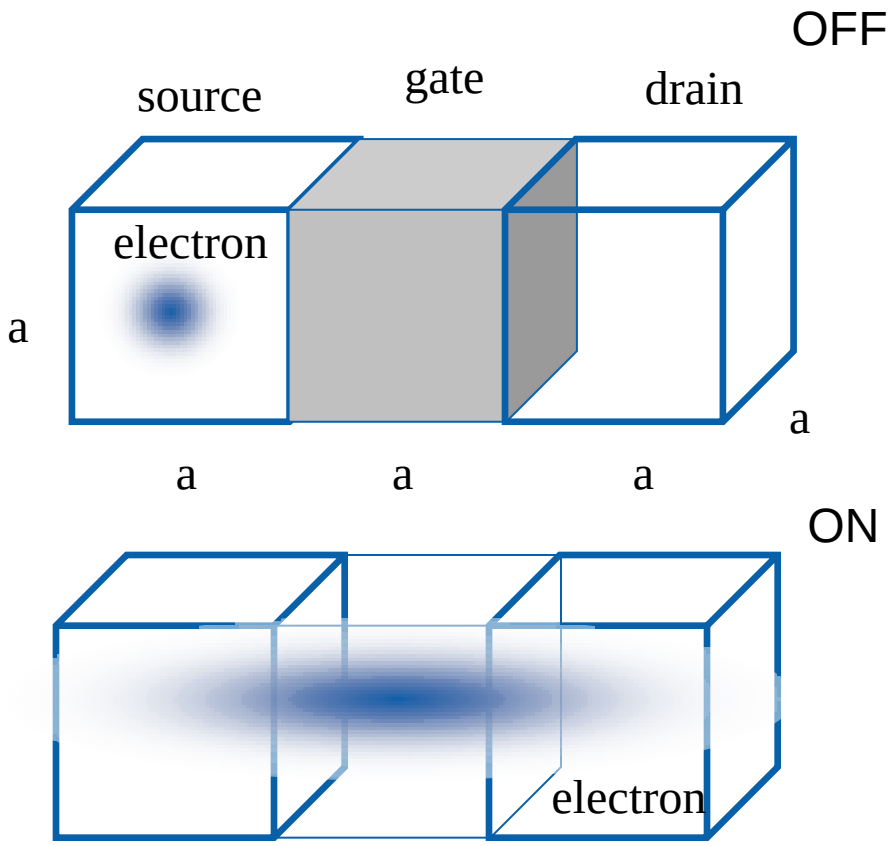
Higher energy = faster switching

$$E_{sw} t_{sw} = \pi\hbar$$

Energy of electron in a transistor is limited by quantum confinement

$$E_{sw} = \frac{3\pi^2\hbar^2}{2ma^2}$$

Gate raised = confined in the source.
Gate lowered = can travel to drain.

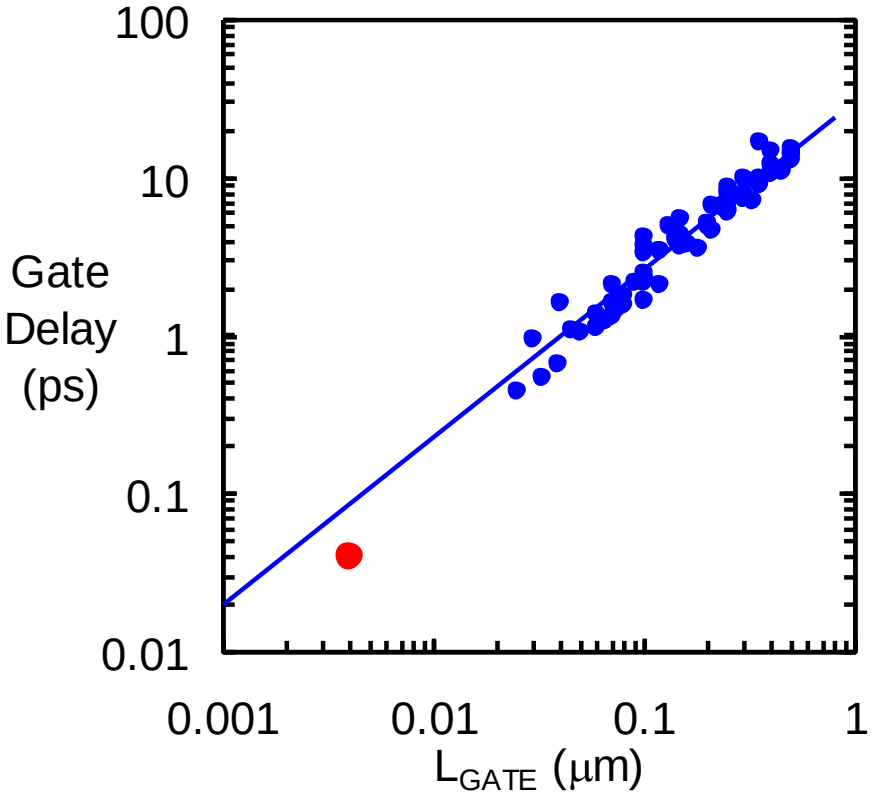


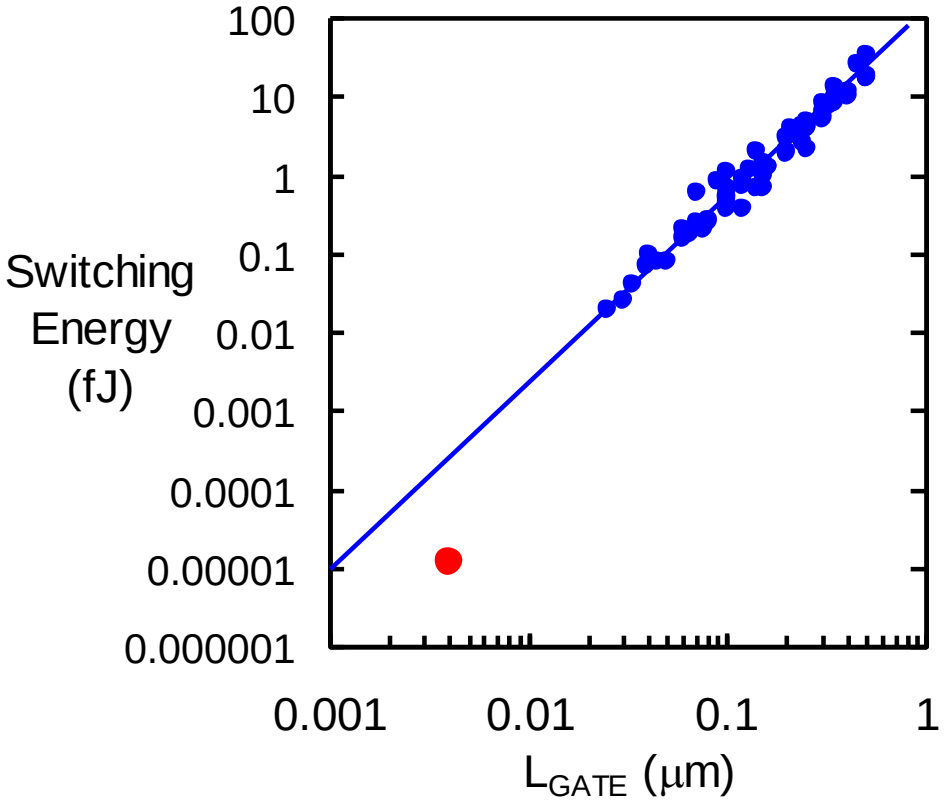
Solving equations together gives limits

$$a = 3.8\text{nm} \qquad t_{sw} = 27\text{fs} \qquad E_{sw} = 1.2\times10^{-20}\,\text{J} = 78\text{meV}$$

* Zhirnov et al., Proc. IEEE 91, 1934 (2003) and Nikonov and Bourianoff, JSNM 21, 497 (2008).

MOSFET Scales Towards the Limit





Current CMOS device scaling close to the ideal limits

* Data courtesy of Robert Chau (Intel)

How long is left for Moore's law

Intel's generation to HVM	
2013	14nm
2017	7nm
2021	3.5nm
ITRS start of production	
2012	32nm
2018	15nm
2024	7.5nm
2030	3.8nm

Scaling might end between 2021 and 2030

But it is NOT the end of Moore's law:

better architectures, 3D circuits.

Summary

- Moore's law = 0.7 size every 2 years
- Despite trends, Intel developers manage to improve performance
- Tri-gate transistors = major advance
- Fundamental laws limit size scaling to ~4nm